

# Statistical Learning Techniques for the Estimation of Lifeline Network Performance and Retrofit Selection

JASON WU	JACK W. BAKER
Thornton Tomasetti	Stanford University
jwu@thorntontomasetti.com	bakerjw@stanford.edu

---

## Abstract

The reliability of water supply networks subjected to catastrophic events is a crucial concern to communities, but our ability to assess these systems is limited by their size and complexity. This paper proposes a statistical learning technique, Random Forests, to efficiently estimate network performance in place of direct physical simulation. This technique uses a set of explanatory metrics that describe the impact of seismic damage to network behavior. The approach is applied to a case study network, the Auxiliary Water Supply System of San Francisco. The resulting statistical model is shown to replicate network performance estimates from flow-based hydraulic simulation, and to provide good performance in identifying components to retrofit to improve the reliability of the system. The favorable performance and computational advantages of this approach make it an attractive tool for infrastructure reliability and risk mitigation analyses.

## 1. Introduction

The reliability and resilience of water supply networks have received much attention due to their vulnerability in extreme events but criticality to community recovery. These networks are complex, and system reliability assessment, while considering a range of possible future disruptive events, is thus computationally challenging.

The reliability of water supply networks may be rigorously evaluated using hydraulic simulation, which estimates the available head, pressure, and flow in the network given a particular (or disrupted) state of the network [e.g. 1, 2, 3, 4, 5], while the resilience of water supply networks may be characterized using a time-varying analysis of the recovery of the network from its disrupted state [e.g. 6, 7, 8, 9, 10, 11]. However, these analytical methods become exceedingly difficult for large networks [12, 13]. Disruptive events further complicate the analysis, as effective network optimization and intervention actions require consideration of a range of disruptive events [8, 14]. Some studies deal with this complexity by focusing on small networks, or using a topological model to assess network performance [e.g. 15, 16, 17, 18], correlating topological properties with reliability and resilience measures [e.g. 19, 20, 3], or formulating a more easily calculable heuristic for reliability and resilience [e.g. 21, 22, 23, 24]. The above strategies have limitations due to the omission of the physical properties and dynamics of the network [12, 6, 25, 26], and so some recent studies aim to incorporate such properties [12, 25, 13]. Rapidly assessing the network performance while considering the network’s physical attributes has two benefits: identifying intervention actions (e.g. retrofitting) to mitigate impacts from potential future disruptive events, and rapidly developing a repair strategy following a disruptive event.

Data-driven analysis has the potential to address the above challenges. Guikema [27] and Perrin [28] discuss the application of statistical learning models to network reliability problems. Han et al. [29] estimate the spatial distribution of power outages during hurricanes with generalized linear models and principal component analysis that use various properties of the landscape, power network components, and hurricane

as explanatory variables. Rokneddin et al. [30] apply Random Forests and Support Vector Machines to estimate the reliability of bridge networks using bridge failure probabilities as predictor variables. Robles-Velasco et al. [31] apply logistic regression and support vector classification to predict pipe failures.

This paper utilizes Random Forests to predict water supply network reliability subjected to earthquakes. The Random Forest model is trained to predict hydraulic network performance, given the damaged state of the network. One key aspect of statistical learning is that the chosen input parameters must sufficiently predict network performance and be easily calculable for large networks. Thus, this work explores the use of a small set of explanatory metrics that reflect the impact of network topology and conveyance of resources due to the damage inflicted by the earthquake. Candidate explanatory metrics may be found in operations research literature regarding the characterization of infrastructure networks [e.g. 32, 33, 34, 35]. Additional metrics aiming to capture the hydraulic impacts of damaged components are formulated in this work to supplement existing metrics from literature. The trained Random Forest model is then utilized to assess the reliability of a case study network and assess the efficacy of identified retrofit strategies in improving its reliability.

This paper is organized as follows. Section 2 describes the Random Forest methodology. Section 3 discusses explanatory metrics and compiles the set of predictor variables to be used for model construction. Section 4 analyzes the predictive power of the Random Forest models by comparing the predictions with those estimated using hydraulic simulation, and evaluates the impact of training data size. Section 5 evaluates the use of the models for retrofit selection. Section 6 concludes by summarizing the findings of this paper and discussing directions for future research.

## 2. Network Performance Estimation

### 2.1. Random Forests

The Random Forest is an ensemble statistical learning method that aggregates a large number of decision trees constructed using bootstrap samples [36]. It is a variant of the tree bagging procedure, in which each tree is grown using its own bootstrap samples as well as a random subset of the predictor variables for training in order to de-correlate trees [37]. The main steps for constructing a Random Forest model are as follows:

1. Compile bootstrap samples from the pool of training data points.
2. Randomly select a subset of predictor variables to be used to construct the tree.
3. Select the predictor variables from the chosen subset to be associated with nodes in the tree, starting with the top node.
4. Recursively split nodes in the tree and distribute the bootstrap samples from that node according to the predictor variable associated with that node, until the minimum number of data samples is achieved at each node.
5. Repeat the above steps a large number of times and aggregate.

Other statistical learning techniques considered in this study were linear regression methods such as Lasso, Support Vector Machine Regression, and Neural Networks. The Random Forest methodology was selected for its ease of implementation, minimal tuning requirements, and speed of model construction and evaluation. Trees are also able to capture complex interactions in the predictor variables and have relatively low bias [e.g. 36, 30, 38].

In this work, each tree is constructed using a randomly selected subset of the predictor variables composed of one-third of the total number of predicted variables and a bootstrapped sample equal to the total number of training data points. Each split (leaf) in the tree incorporates a minimum of five data samples. These values for the hyperparameters of the Random Forest models are the suggested values found in [39, 37]. Each Random Forest model grows a maximum of 500 regression trees. This work uses the Random Forest implementation in the Statistics Toolbox of Matlab, named the `Treebagger` class [39].

The Random Forest can estimate the relative importance of its predictor variables using out-of-bag permuted variable delta error (OOBPVDE) [39]. For each regression tree, there are “out-of-bag samples”

omitted from training. For each tree grown in the model, the value of a particular predictor variable is randomly permuted using the values in the out-of-bag samples, and the resulting prediction error is recorded. The increase in prediction error is averaged over all trees and divided by the standard deviation across the entire ensemble. This value, the OOBPVDE associated with that variable, describes its contribution to the predictive power of the model. A higher OOBPVDE indicates higher relative importance.

## 2.2. Example network

To illustrate the Random Forest strategy, the Auxiliary Water Supply System (AWSS) of San Francisco is used as an example network. This high-pressure network provides water for firefighting, and has been the subject of many past studies [e.g. 40, 41]. Additionally, a major retrofit project is underway to improve its seismic reliability [42]. The AWSS model is composed of 6,261 nodes, including two tanks, one reservoir, and two pump stations, and 6,307 pipe segments spanning 205 kilometers, as shown in Figure 1. Topologically, the northeast portion of the network is relatively dense, with high redundancy of pipelines, whereas the southern and western part of the network is relatively sparse. Hydraulically, the northeast portion of the network has a higher number of nodes requiring water and higher quantity of water demanded. The AWSS functions independently from other networks, thereby minimizing network interdependency effects on its performance.

To analyze this network, a comprehensive set of earthquake scenarios, with incremental magnitudes, rupture distances, locations, etc., is generated from [43], and peak ground velocities for each scenario are calculated using the Boore and Atkinson [44] ground motion model according to methods described by [2, 45]. 1,820 ground motion realizations are generated from these methods, each with an associated probability of occurrence to maintain the hazard consistency. This set of ground motion realizations is combined with pipe damage probabilities from Jeon and O’Rourke [46] to produce realizations of pipe damage in the network. For simplicity, nodes are not susceptible to damage in this work. The damaged networks are subjected to nodal demands for firefighting water, as documented in [47]. To facilitate computation, pipelines are serially aggregated in the model: pipe segments in series are modeled as a single pipeline, resulting in a smaller network that contains no nodes connected to exactly two pipe segments.

## 2.3. Network performance estimation

The performance of the network is calculated using hydraulic flow simulation [48, 49], and quantified by the nodal unsatisfaction. Nodal unsatisfaction is defined to be the proportion of nodes in the network that do not have their water demands satisfied. This is a continuous variable that takes a value of 0 when no nodes have demands satisfied and 1 when all are satisfied. This nodal unsatisfaction is calculated for each of the previously described 1,820 simulations.

These simulations are used to train the Random Forest model to predict nodal unsatisfaction, and a second set of simulations are used to evaluate the prediction error of the resulting model. This paper explores the use of two different sets of predictor variables to evaluate their efficacy in prediction:

The first approach is to use the set of damaged components as predictor variables for the Random Forest, denoted here as the *Naïve strategy*. This strategy directly links the performance of network components to the performance of the network, and, with sufficient training and refinement of the model, would abstract away the complexities of network operation. For the AWSS, there are 6,307 pipe segments, and thus 6,307 predictor variables for this strategy. Each variable may take an integer value of 0 or 1, indicating no damage or damage, respectively. One issue with the Naïve strategy is the large parameter space of the predictor variables. If the amount of training data is small relative to the parameter space, then the resulting model may not adequately accommodate possible future observations. Additionally, when there are a large number of predictor variables, but a relatively small number of important or relevant variables, those variables are less likely to be chosen to construct each tree, resulting in a poorly performing Random Forest model [36]. As the Naïve strategy uses the network component states as the predictor variables, a large proportion of components may have a relatively small impact on network performance, especially for large networks.

The second approach is to first translate the network component states into a smaller set of explanatory network metrics, which summarize the impact of damage to the network. This smaller set of metrics are then

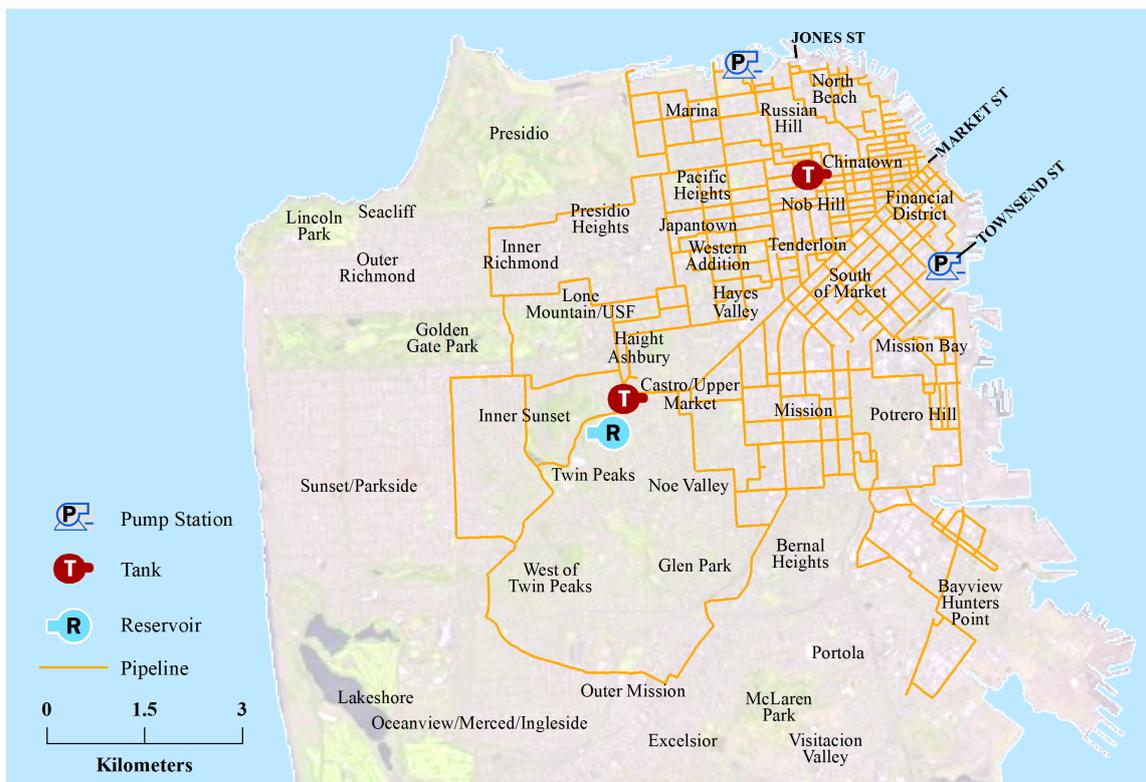


Figure 1: Major components of the Auxiliary Water Supply System in San Francisco.

used as the predictor variables for Random Forest training and forward prediction. This method reduces the parameter space of the predictor variables and provides predictor variables that are more strongly linked with network performance. This method is denoted here as the *Metric strategy*. The next section discusses the explanatory metrics for the Metric strategy.

### 3. Candidate Explanatory Metrics

This section discusses the considered explanatory metrics. Some of the metrics are popular in the network literature, and others were formulated here based on domain-specific features of water networks. The importance of each variable is assessed, and a method proposed to optimize the set of predictors used for statistical learning.

#### 3.1. Metric Definitions

The 23 explanatory metrics utilized in this paper contribute 49 predictor variables to train the Random Forest model, and are summarized in Table 1. The Random Forest model constructed using these 49 predictor variables is denoted as the *Full Set* and used for comparative purposes later in this paper. The following subsections define the three metrics that are determined in the analysis below to be most useful. The remaining metrics are defined in the electronic supplement.

##### 3.1.1. Spectral Radius

The spectral radius, denoted  $r$ , is a spectral measure conceived to capture the robustness of networks against the spread of viruses [32], and describes the interconnectivity between nodes. The spectral radius is calculated as follows.

$$r = \max_{1 \leq i \leq N} |\lambda_i| \quad (1)$$

where  $\lambda_i$  is the  $i^{th}$  eigenvalue of the adjacency matrix  $A$ . For the AWSS, this metric’s value varies between 3.12 and 3.38 across the data points.

##### 3.1.2. Proportion of Affected Tiers

The proportion of affected tiers, denoted  $pat$ , is an experimental topological measure formulated in this paper. It is defined as the proportion of links that are downstream of damaged links according to topological sort. This is computed by finding the damaged link with the highest tier from each source node, extracting the sets of links with lower tiers than that damaged link, and amalgamating these sets across all source nodes. This metric aims to capture the impact of damaged links on the distribution of flow. It is calculated as follows.

$$pat = \frac{|\bigcup_{i \in S} lower_{link}^i|}{m} \quad (2)$$

where  $S$  is the set of all source nodes,  $lower_{link}^i$  indicates the set of all links downstream from damaged links according to a topological sort from source node  $i$ ,  $m$  is the number of links in the network, and  $|\bullet|$  describes the cardinality of the set  $\bullet$ . A value of 0 indicates that there are no damaged pipes or the damaged pipe is the lowest tier from all reachable sources, and a value of 1 indicates that all pipes connected to source nodes are damaged. For the AWSS, this metric varies between 0 and 0.724 across the data points.

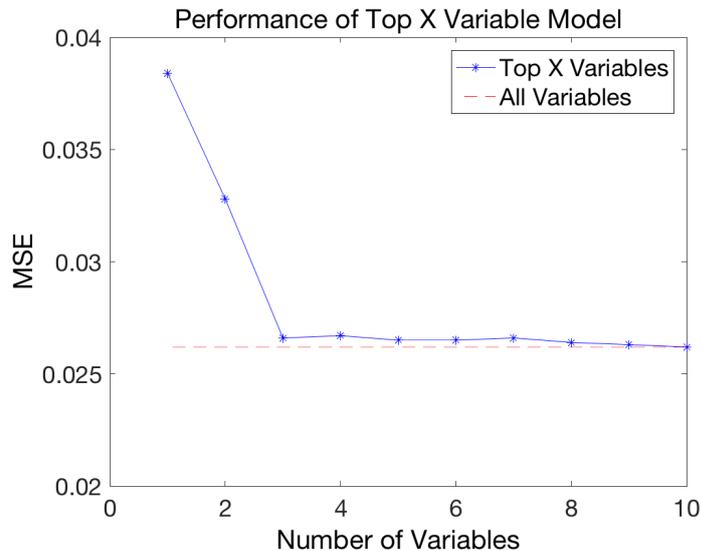


Figure 2: Mean squared error (MSE) of models constructed using a subset of the most important explanatory variables, described by the solid blue line. These values are compared to the mean squared error of the model using all variables, indicated by the dashed red line.

### 3.1.3. Critical Fraction

The critical fraction, denoted  $f_c$ , is a topological measure that describes the theoretical value for the proportion of nodes which need to be removed to completely destroy the largest cluster in a network [50]. It is calculated as follows.

$$f_c = 1 - \frac{1}{\frac{\langle k_i^2 \rangle_i}{\langle k_i \rangle_i} - 1} \quad (3)$$

where  $k_i$  is the degree of node  $i$  and  $\langle \bullet \rangle_i$  denotes the averaging of the values  $\bullet$  over all nodes  $i$ . For the AWSS, the serial aggregated network is used to facilitate the calculation of this metric. This metric varies between 0.472 and 0.508 across the AWSS data points.

### 3.2. Importance of Predictor Variables

The importance of each variable may be estimated by observing its out-of-bag permuted variable delta error (OOBPVDE) [39]. An initial Random Forest model is constructed using 1,820 training data points, and the OOBPVDE is calculated using the resulting model. Table 1 provides the importance ranking of the predictor variables using OOBPVDE.

Using this importance ranking, an incremental analysis is performed. First, a Random Forest is constructed with only the most important variable, and its prediction error is evaluated. Subsequent models incrementally add the next most important variable to the predictor set. Figure 2 shows the mean squared prediction error (MSE) on a test data set for each of these incremental models, compared to the prediction error of a model that uses all predictor variables.

$$MSE = \langle [x_j - x'_j]^2 \rangle_j \quad (4)$$

where  $x_j$  is the true value of the nodal unsatisfaction for data point  $j$ ,  $x'_j$  is the predicted value for the nodal unsatisfaction for data point  $j$ , and  $\langle \bullet \rangle_j$  denotes the average of the quantities  $\bullet$  for all data points  $j$ . Further explanation of the test data set will be presented in Section 4

Observing Figure 2, the three most important predictor variables (spectral radius, proportion of affected tiers, and critical fraction) provide predictive power comparable to the full set of variables. Spectral radius

and critical fraction quantify the interconnectivity between nodes (i.e. redundancy of water flow paths), while the proportion of affected tiers quantifies the number of flow paths impacted by damage. Thus, these metrics indicate that path redundancy (or the loss thereof) is closely associated with nodal dissatisfaction. A Random Forest model using these three predictor variables, termed the *Optimized Set*, is evaluated further below. While other effective combinations of predictor variables may exist, Figure 2 indicates that improvements to predictive power would be minimal.

Table 1: Predictor variables from the explanatory metrics explored in this work.

Variable Name	Symbol	Importance Ranking
Average Degree	$k$	11
Meshedness Coefficient	$\alpha$	14
Link Density (Network Density)	$\rho$	10
Critical Fraction	$f_c$	3
Closeness (source 1)	$C_1$	29
Closeness (source 2)	$C_2$	33
Closeness (source 3)	$C_3$	31
Closeness (source 4)	$C_4$	23
Closeness (source 5)	$C_5$	19
Closeness (source 6)	$C_6$	28
Closeness (source 7)	$C_7$	24
Directed Closeness (source 1)	$C_1^D$	9
Directed Closeness (source 2)	$C_2^D$	46
Directed Closeness (source 3)	$C_3^D$	44
Directed Closeness (source 4)	$C_4^D$	36
Directed Closeness (source 5)	$C_5^D$	42
Directed Closeness (source 6)	$C_6^D$	45
Directed Closeness (source 7)	$C_7^D$	47
Average Shortest Path Length (source 1)	$L_1$	30
Average Shortest Path Length (source 2)	$L_2$	34
Average Shortest Path Length (source 3)	$L_3$	22
Average Shortest Path Length (source 4)	$L_4$	26
Average Shortest Path Length (source 5)	$L_5$	18
Average Shortest Path Length (source 6)	$L_6$	27
Average Shortest Path Length (source 7)	$L_7$	32
Directed Average Shortest Path Length (source 1)	$L_1^D$	13
Directed Average Shortest Path Length (source 2)	$L_2^D$	48
Directed Average Shortest Path Length (source 3)	$L_3^D$	41
Directed Average Shortest Path Length (source 4)	$L_4^D$	35
Directed Average Shortest Path Length (source 5)	$L_5^D$	40
Directed Average Shortest Path Length (source 6)	$L_6^D$	43
Directed Average Shortest Path Length (source 7)	$L_7^D$	49
Algebraic Connectivity	$\mu_{n-1}$	20
Spectral Gap	$\Delta$	15
Spectral Radius	$r$	1
Connectivity Loss	$C_L$	39
Reachability Loss	$R_L$	37
Redirection	$rd$	38
Flow Loss Score (Shortest Path)	$fl_{SP}$	7
Flow Loss Score (Betweenness)	$fl_B$	21
Flow Loss Score (Topological Sort)	$fl_{TS}$	17
Proportion of Affected Tiers	$pat$	2
Average Proportion of Affected Demand	$apad$	4

Continued on next page

Table 1 – continued from previous page

Variable Name	Symbol	Importance Ranking
Overall Proportion of Affected Demand (Top. Sort)	$opad_{TS}$	16
Path Reduction	$pr$	25
Weighted Path Reduction (Upstream Flow)	$wpr_{UF}$	12
Weighted Path Reduction (Downstream Links)	$wpr_{DL}$	6
Weighted Path Reduction (Downstream Demand)	$wpr_{DD}$	5
Overall Proportion of Affected Demand (DirectedPaths)	$opad_{DP}$	8

#### 4. Predictive Power of Random Forest Models

This section assesses the predictive power of the developed Random Forest models. The first subsection discusses the predictive power of the three primary considered models. Each of these models are trained with 1,820 data points generated as described in Section 2.2 and tested with a second set of 1,820 realizations, referred to henceforth as the test data. The second subsection studies the predictive power of the Random Forest models as a function of the amount of training data used.

##### 4.1. Performance of Primary Random Forest Models

Figure 3 shows scatter plots of 1,820 data points from the test data describing the relationship between the predicted nodal unsatisfaction versus the true value, colored based on point density. Blue points signify a sparse concentration of points in the vicinity, while red point signify a dense concentration. The solid line indicates the values at which the predicted and true values are equal. The three subfigures depict the results for the Naïve strategy, Metric strategy with the Full Set of variables, and Metric strategy with the Optimized Set of variables.

In Figure 3a, predicted values vary between 0.35 and 0.8, resulting in a large disparity between the true and predicted values in the dense regions where the true value is around 0.2 and 1.0. This reflects the difficulty that Random Forest models have in extrapolating. Due to the large parameter space, it is nearly certain that the test data contain pipe damage combinations that were not observed in the training data, and with this model parameterization the Random Forest struggles to make predictions in such cases. Nevertheless, there is some correlation between the true and predicted values.

In Figure 3b, there is a stronger correlation between the predicted and true values as compared to the Naïve strategy. There is a dense region near the red line around the value of 0.2 and 0.7. The model underestimates nodal unsatisfaction when the true value is near 1.0, though to a lesser degree than the Naïve strategy. Figure 3c closely resembles that of Figure 3b. Thus, for this network, the Optimized Set of variables can efficiently predict response.

To supplement Figure 3, the predictive power is quantitatively evaluated by calculating the mean squared error (MSE) and mean absolute error (MAE) of the predictions on the test data, where MSE was defined above in Equation 4 and MAE is defined as

$$MAE = \langle |x_j - x'_j| \rangle_j \quad (5)$$

where  $|\bullet|$  denotes an absolute value and other variables are as defined in Equation 4.

For the Naïve strategy, Metric strategy with Full Set, and Metric strategy with Optimized Set, the MSE over all data points is 0.0343, 0.0260, and 0.0266, respectively, and the MAE over all data points is 0.1523, 0.1248, and 0.1251, respectively. Thus, there is a perceivable disparity in overall predictive power between the Naïve strategy and Metric strategies, while the Optimized Set nearly replicates the power of the Full Set.

The MSE and MAE values for bins of true values are shown in Figure 4. The Naïve model performs relatively poorly for nodal unsatisfaction values below 0.3 and above 0.7, which is consistent with the

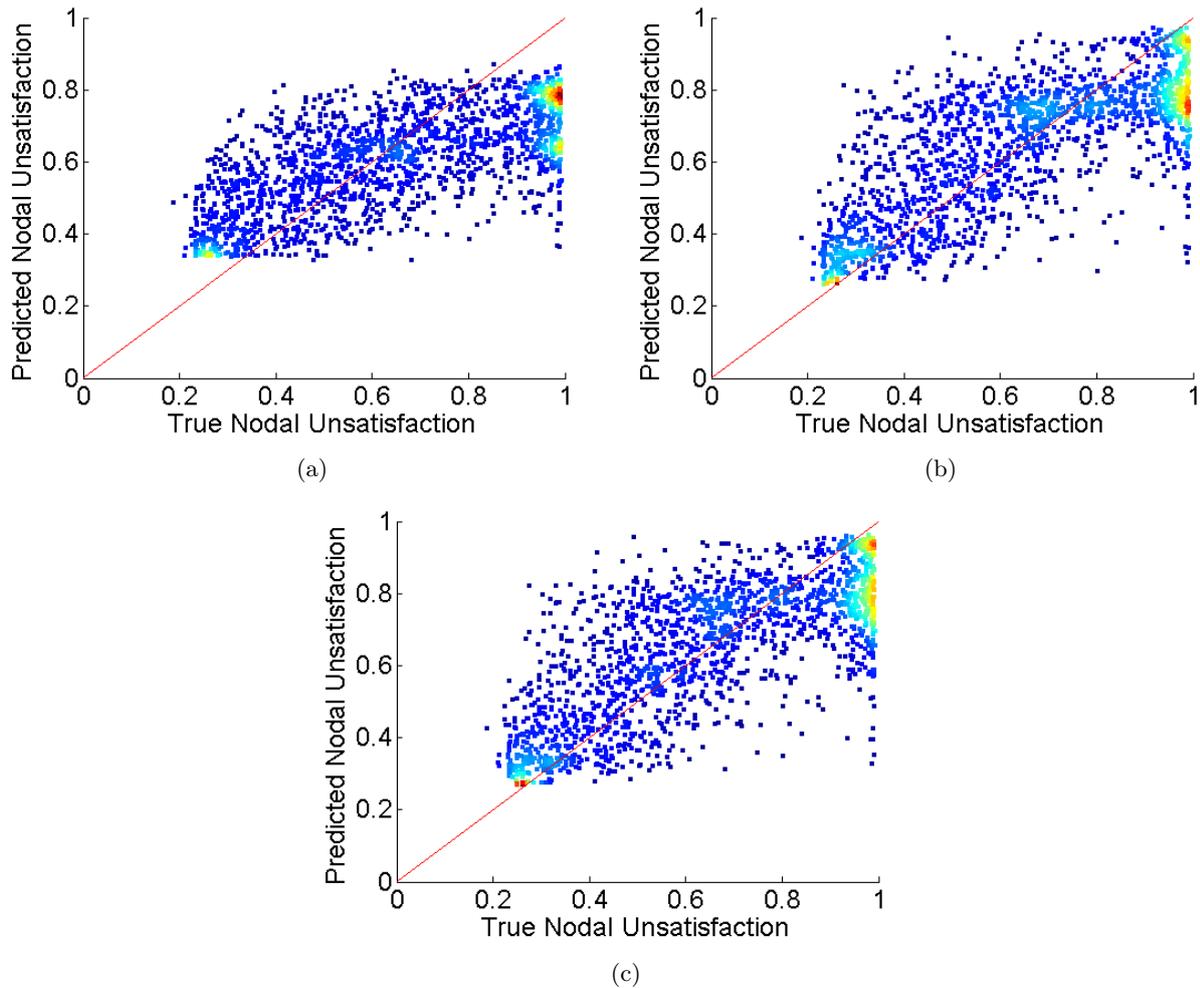


Figure 3: Scatter plots depicting the true nodal unsatisfaction from the hydraulic model versus the predicted nodal unsatisfaction from the Random Forest model following (a) the Naïve strategy, (b) the Metric strategy, with the Full Set of variables, and (c) the Metric strategy, with the Optimized Set of variables. Each point on the plot represents one simulation for a damaged network. The coloration signifies the density of the point cloud: blue points signify low density of points in the vicinity, and red points signify high density of points in the vicinity.

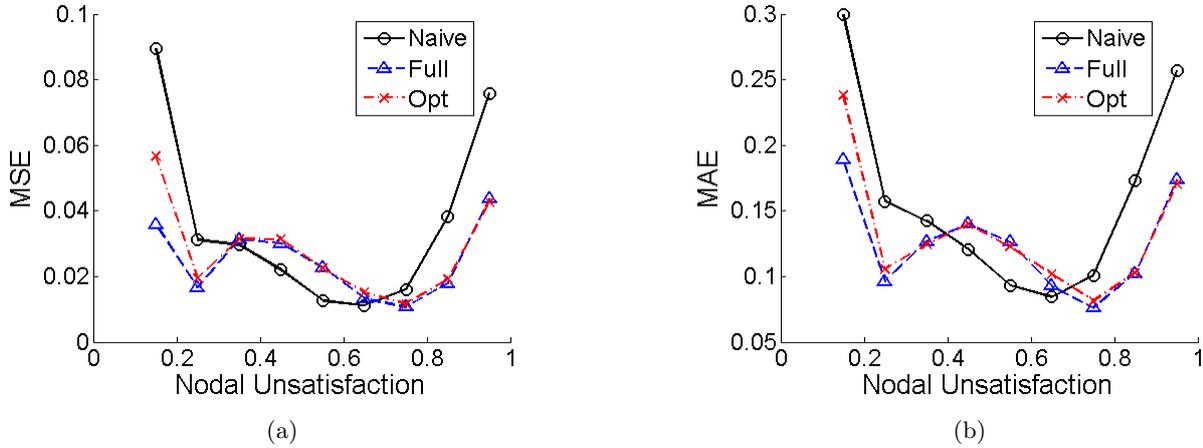


Figure 4: (a) Binned mean squared error (MSE), and (b) binned mean absolute error (MAE) of the nodal unsatisfaction using the Naive and Metric models..

observations of Figure 3a. For the two Metric strategies, the Optimal Set is inferior only in the regions of low nodal unsatisfaction values, which are not as important since the focus is on highly disrupted networks. These results further demonstrate the good performance of the Optimized Set Random Forest model.

#### 4.2. Sensitivity of Predictive Power to Training Data Size

The previous subsection demonstrated the capability of the Metric strategy, Optimized Set to maintain superior predictive power despite the use of only three predictor variables. Recall that these models have been constructed using 1,820 training data points. This subsection aims to push the efficiency of this strategy further by reducing the size of the training data and observing the impact to predictive power.

Next models are built using the Metric strategy with the Optimized Set of predictor parameters, but using differing amounts of training data. The first model (designated *Tr500*) uses 500 training data points. The second model (*Tr100*) uses 100 training data points. The third model (*Tr10*) uses only 10 training data points, to observe the effects of very little model training. All three models use subsets of the 1,820 data points originally used for training.

Figure 5 presents the scatter plots of the predicted values versus the true values for these models. *Tr500* performs well and resembles the performance of the previous model using 1,820 training data points. Meanwhile, *Tr100* appears to exhibit a constraint in predicted values below 0.5, and *Tr10* exhibits a strong constraint in only predicting values between 0.6 and 0.8, reminiscent of the scatter plot for the Naïve model. This again demonstrates the limitation of the Random Forest methodology to extrapolate if the training is inadequate. On the other hand, for *Tr100* the region with higher nodal unsatisfaction appears to perform nearly as well as the other models. For *Tr10*, there is still a slight correlation between the predicted and true values, as the bottom left and top right areas of the scatter plot exhibits a high density of data points. The predictive power of the Random Forest has severely deteriorated, however, due to the extremely limited training data.

To supplement this analysis, the MSE and MAE of these models are calculated. The overall MSE of *Tr500*, *Tr100*, and *Tr10* are 0.0274, 0.0356, and 0.0496, respectively, and the overall MAE values are 0.1266, 0.1409, and 0.1939, respectively. Recall that the MSE and MAE for *Opt* are 0.0266 and 0.1259. The binned MSE and MAE for these models are also calculated, and shown in Figure 6.

Consistent with the previous analysis, *Tr500* adequately replicates the performance of *Opt*, and *Tr100* performs well for larger nodal unsatisfaction values. *Tr10* performs well within a limited region, though this is an artifact of its constrained predictions and not an indicator of predictive power. These results imply that 500 training data points may sufficiently cover the parameter space to yield adequate predictions.

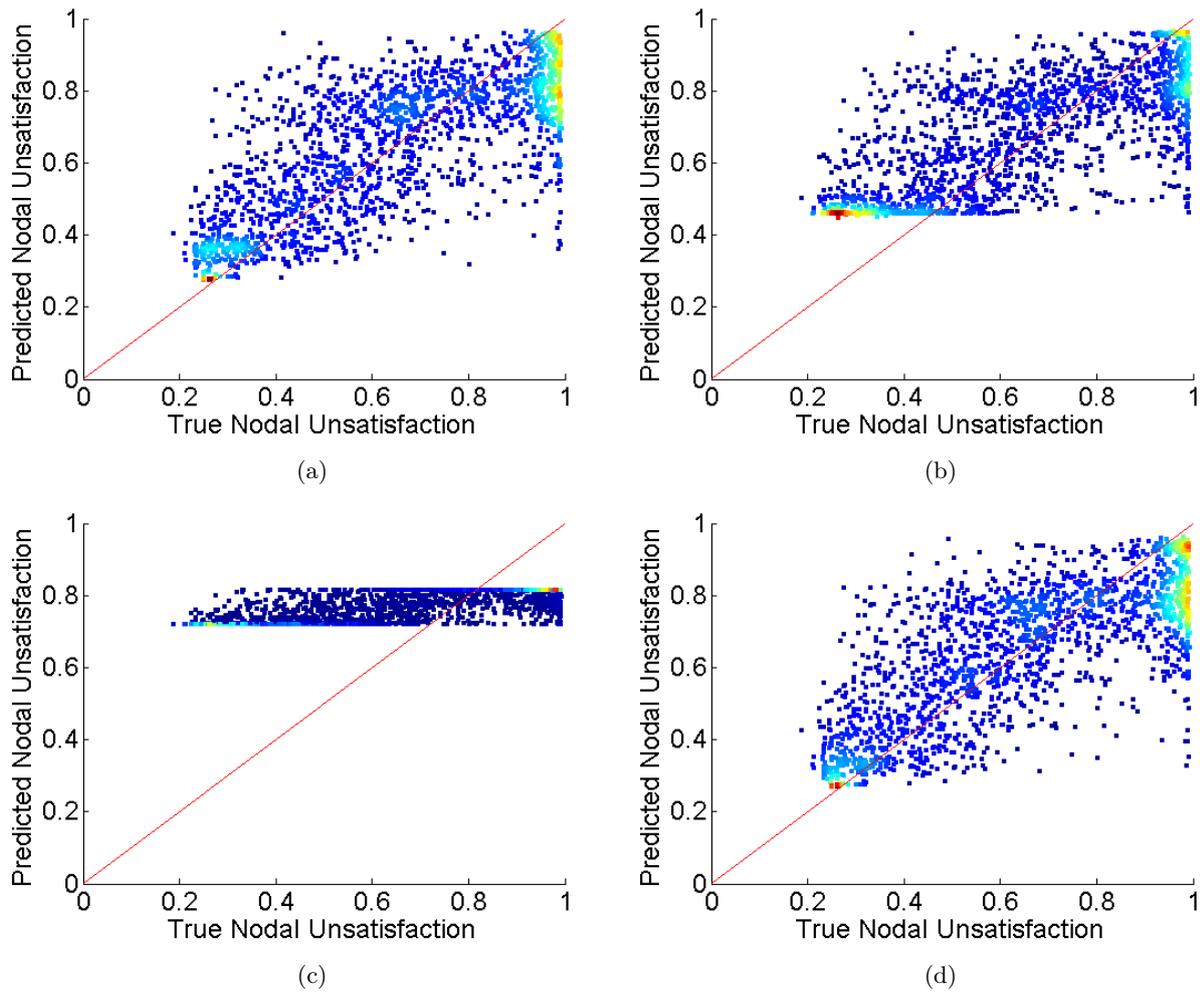


Figure 5: Scatter plots depicting the true nodal unsatisfaction from the hydraulic model versus the predicted nodal unsatisfaction from the Random Forest models using the Metric strategy, Optimized set, but differing amounts of training data. (a) Using 500 training data points (*Tr500*). (b) Using 100 training data points (*Tr100*). (c) Using 10 training data points (*Tr10*). (d) Using 1,820 data points (*Opt*).

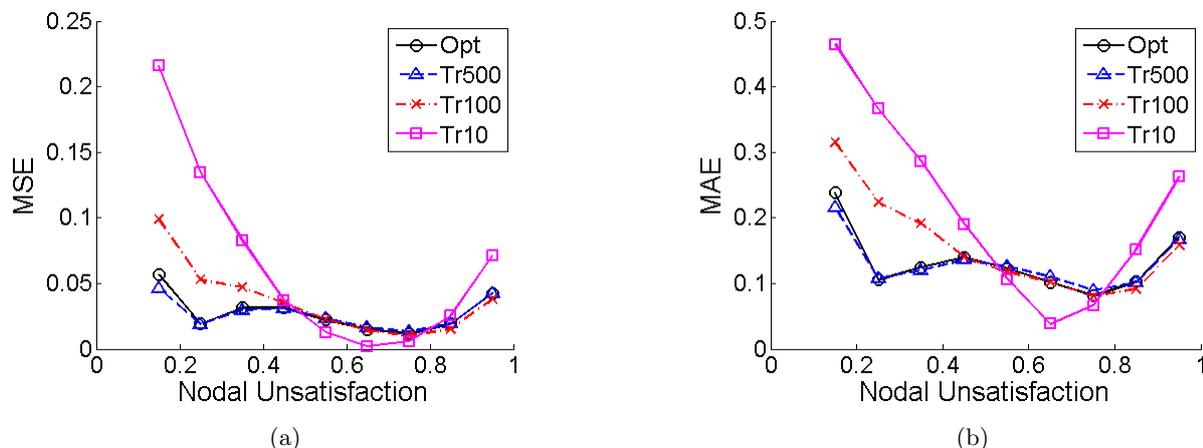


Figure 6: (a) Binned mean squared error (MSE), and (b) binned mean absolute error (MAE) using the Metric models with varying amounts of training data.

Additionally, a training data size of 100 may still perform adequately if only high nodal unsatisfaction is of interest.

## 5. Statistical Learning Strategies for Retrofit Selection

While Section 4 assessed the performance of the proposed statistical models in terms of their predictive power, this section evaluates the performance of the models in their application to network component retrofit selection. The Random Forest models are used to predict network performance, and those simulations yielding high nodal unsatisfaction are used for pipe segment retrofit selection using methods described in [1]. These results are compared with those using hydraulic simulation to determine network performance.

### 5.1. Effectiveness of Retrofits Using the Optimized Random Forest Model

This section first presents the effectiveness of retrofits selected using simulations with nodal unsatisfaction above the threshold of 0.8. The Random Forest model using the Metric strategy, with the Optimized Set of variables, is used to predict the nodal unsatisfaction values. This model is trained using 1,820 data points, and used to predict damage simulations to be above or below a nodal satisfaction of 0.8. 55 km of pipeline, corresponding to roughly 25% of the pipelines, are selected for retrofit according to [1]. Here, retrofits are assumed here to make pipes invulnerable to damage. Figure 7 presents the network reliability after the application of retrofits in the form of the annual exceedance rate of nodal unsatisfaction. The network with retrofits identified using the Random Forest model is labeled as *Opt*, and this is compared to the network with retrofits using hydraulic simulation, labeled as *Hydraulic*.

Figure 7 indicates that the Random Forest model predictions are capable of replicating the retrofit effectiveness of those calculated using hydraulic simulation. Additionally, the retrofits selected by the Random Forest model and the hydraulic simulation model share approximately 95% of pipeline, by length. So, despite the variability in predicted values observed in the scatter plots, the Random Forest models perform well in selecting effective retrofits.

Note that despite the moderate prediction error indicated by the scatter plot in Figure 3 and the binned MSE and MAE in Figure 4, the Random Forest model largely matched the hydraulic model in identifying important pipes. However, the application of these models to retrofit selection changes the nature of the prediction. Since simulations are labeled as “highly disrupted,” the objective becomes the classification of maps above or below the threshold. The model need only assign the correct label, rather than precisely predict nodal unsatisfaction.

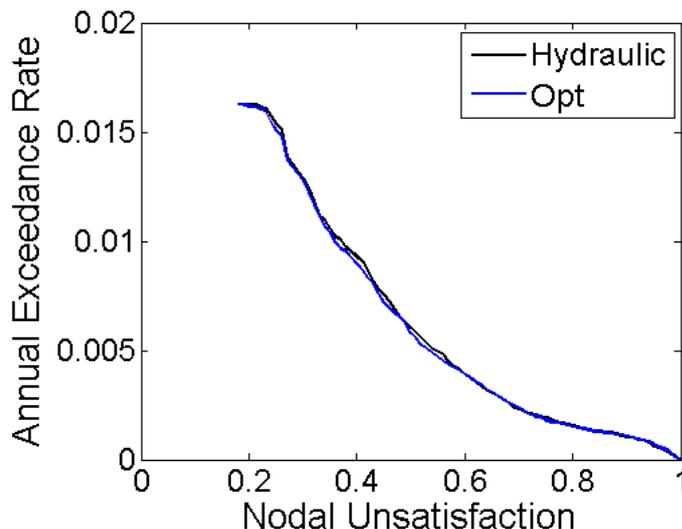


Figure 7: Network performance in terms of the annual exceedance rate of the nodal unsatisfaction after the application of retrofits. Retrofits are selected using either hydraulic simulation (labeled Hydraulic) or a Random Forest model using the Metric strategy, with the Optimized Set of variables (labeled Opt).

The power of the Random Forest models to classify is assessed using the confusion matrix. The confusion matrix describes the proportion of data that is correctly or incorrectly labeled. Given two classification labels  $A$  and  $\sim A$  (not  $A$ ), the confusion matrix is composed of the *true negative* (the proportion of data correctly labeled as  $\sim A$ ), *false positive* (the proportion of the data incorrectly labeled as  $A$ ), *false negative* (the proportion of the data incorrectly labeled as  $\sim A$ ), and *true positive* (the proportion of the data correctly labeled as  $A$ ). Using the confusion matrix, we may determine the classification accuracy—the proportion of data that is correctly labeled—by summing the true negative and true positive proportions. Furthermore, as retrofit selection uses those maps that are classified to be above the threshold of 0.8, we also consider the proportion of confounding data—the proportion of the data included for retrofit selection that should not have been included. The proportion of confounding data is the ratio of false positive labels to the sum of the false positive and true positive labels. The proportion of confounding data captures the amount of incorrect data (or noise) integrated into retrofit selection, which may impact the effectiveness of the resulting retrofits.

For the model used in Figure 7, this may be seen in Figure 8, which depicts the scatter plot from Figure 3c with dashed black lines indicating the threshold used to indicate high disruption. Points to the left of the vertical dashed black line are truly below the threshold and should not be included in the retrofit selection procedure. Those right of the line are truly above the threshold should be included in the selection. Those below the horizontal dashed black line are predicted to be below the threshold and are not included in the selection. Those above the line are predicted to be above the threshold and included in the selection. The quadrants formed by these dashed black lines correspond to the elements of the confusion matrix (e.g., the top right quadrant are true positive events). The proportions of the true negative, false positive, false negative, and true positive are 0.62, 0.06, 0.17, and 0.15, respectively. Using these values, the classification accuracy is 0.77, and the proportion of confounding data is 0.31.

The selection of the retrofits using this model suggests that this classification accuracy is high enough, or alternatively the proportion of confounding data is low enough, such to yield an adequate set of simulations to efficiently select retrofits. Despite this moderate proportion of confounding data, the Random Forest model is capable of matching 95% of pipeline retrofits identified using hydraulic modeling. One explanation is that the confounding data does identify a subset of truly important pipelines, and this is reinforced by the contribution of the correctly labeled data. Furthermore, the procedure for retrofit selection is reasonably robust as the threshold of 0.8 is not a strict one—similar network performance was observed when this

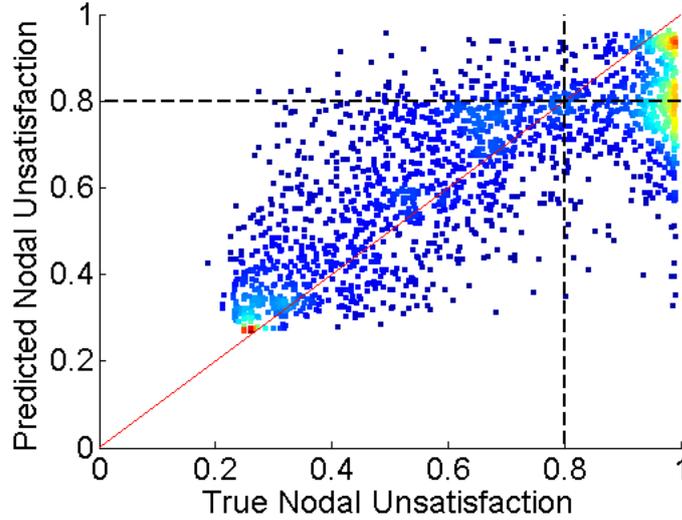


Figure 8: Scatter plot depicting the predictive power of the Random Forest model using the Optimized set of variables. The dashed black lines indicate the threshold to determine high network disruption and define the boundary between class labels of data points (i.e. included or not included in retrofit selection) for the associated classification problem.

threshold was varied [2]. Thus, the random forest models are not required to correctly classify each damage simulation as long as they can identify highly disrupted networks.

### 5.2. Sensitivity of Retrofit Selection Effectiveness to Predictive Power

This subsection performs a sensitivity analysis of the adequacy of the threshold for determining high network disruption and effectiveness of the resulting retrofit selection to the predictive power of the Random Forest models. First, the impact on retrofit effectiveness due to the reduction of training data for the Random Forest model is evaluated. The models corresponding to *Tr500*, *Tr100*, and *Tr10* defined in Section 4 are employed here for retrofit selection. Figure 9 depicts the resulting network performance using these models, compared to the *Hydraulic* and *Opt* curves from Figure 7. An additional curve depicting the network performance with no retrofits is added for comparison and labeled *Baseline*.

Observing Figure 9, the *Tr500* and *Tr100* models appear to perform just as well as *Opt* and *Hydraulic*, while the *Tr10* model exhibits a large disparity in the effectiveness of the retrofits. This is consistent with the previous analysis of the predictive power of these models, in which *Tr500* performed as well as *Opt*, while *Tr10* exhibited perceivable inadequacies. In fact, recalling Figure 5c, the *Tr10* model cannot predict values above 0.8, so the selected retrofits are completely random. Also recall, the *Tr100* model had higher overall prediction errors, but performs well in predicting the higher nodal unsatisfaction cases important for retrofit selection. This further demonstrates the capability of the Random Forest model to perform well with a relatively small amount of training: on the order of 500 data points, and potentially down to 100 data points, depending on the application and the features of the training data.

Next, this section assesses the impact of using the Naïve strategy for nodal unsatisfaction predictions during retrofit selection. Figure 10 depicts the network performance after retrofits are selected and implemented using hydraulic simulation, the Metric model with the Optimized Set, and the Naïve model to determine highly disrupted simulations, whose curves are denoted *Hydraulic*, *Opt*, and *Naive*, respectively.

In Figure 10, interestingly, both Random Forest models appear to perform comparably to the hydraulic simulation. This is despite the disparity in predictive power between the Naïve and Metric strategies. Additionally, the retrofit selection from the Naïve strategy matches approximately 92% of pipeline, by length, with the retrofit selection from the use of hydraulic simulation.

An analysis of the confusion matrix may explain the performance of the Naïve strategy. Its true negative, false positive, false negative, and true positive proportions are 0.676, 0.012, 0.269, and 0.043, respectively.

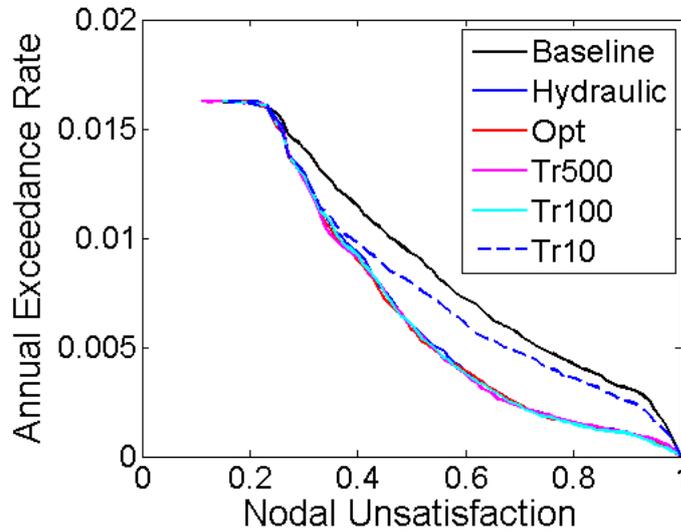


Figure 9: Network performance in terms of the annual exceedance rate of the nodal unsatisfaction after the application of retrofits. Retrofits are selected using either hydraulic simulation (labeled Hydraulic), a Random Forest model using the Metric strategy with the Optimized Set of variables and 1,820 training data (labeled Opt), a model constructed with 500 training data (labeled Tr500), a model constructed with 100 training data (labeled Tr100), and a model constructed with 10 training data (labeled Tr10). The curve labeled Baseline describes the network performance with no retrofits.

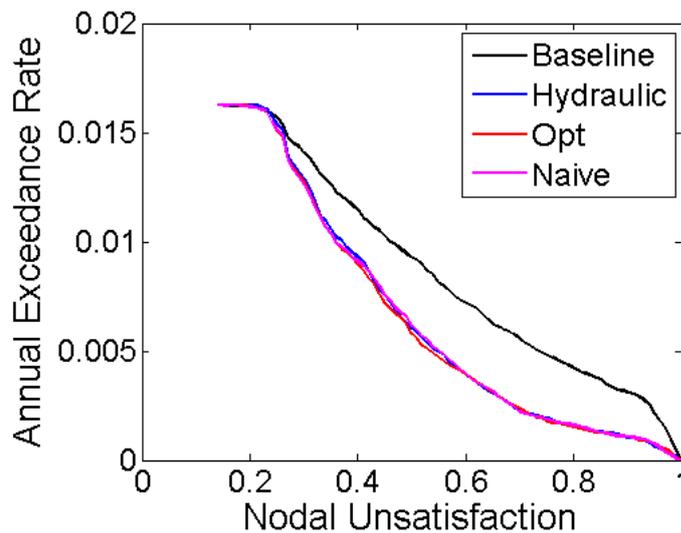


Figure 10: Annual exceedance rate of the nodal unsatisfaction after the application of retrofits. Retrofits are selected using hydraulic simulation (labeled Hydraulic), a Random Forest model using the Metric strategy, with the Optimized Set of variables (labeled Opt), and a Random Forest model using the Naïve strategy. The curve labeled Baseline describes the network performance with no retrofits.

Table 2: Approximate Computation Times of Network Performance Estimation Methods To Simulate or Predict Future Data.

<b>Procedure</b>	<b>Computation Time (seconds)</b>
<b>Hydraulic Simulation</b>	
Simulation, per data point	1,000
<b>Total time (1 Simulation)</b>	<b>1,000</b>
<b>Total time (1,820 Simulations)</b>	<b>1,800,000</b>
<b>Naïve Strategy</b>	
Data assembly and prediction, per data point	< 1
Model training (1,820 training data points)	60,000
<b>Total time (1 Prediction)</b>	<b>60,000</b>
<b>Total time (1,820 Predictions)</b>	<b>60,000</b>
<b>Metric Strategy, Full Set</b>	
Precalculations, supplementary tools	450
Metric calculations, per data point	6
Model training (1,820 data points)	60
Prediction, per data point	< 1
<b>Total time (1 Prediction)</b>	<b>520</b>
<b>Total time (1,820 Predictions)</b>	<b>12,000</b>
<b>Metric Strategy, Optimized Set</b>	
Precalculations, supplementary tools	60
Metric calculations, per data point	1.5
Model training (1,820 data points)	10
Prediction, per data point	< 1
<b>Total time (1 Prediction)</b>	<b>75</b>
<b>Total time (1,820 Predictions)</b>	<b>3,000</b>

This yields a classification accuracy of 0.72 and a proportion of confounding data of 0.21. Thus, although the Naïve strategy is poorer at predicting nodal dissatisfaction and poorer at classifying maps with high nodal dissatisfaction, it has a competitive classification accuracy (compared to 0.77 for the Metric strategy, Optimized Set) and a lower proportion of confounding data (compared to .31 for the Metric strategy, Optimized Set). Thus, while the Naïve strategy is less accurate overall, the data used during retrofit selection is just as accurate by proportion, which may explain its ability to select efficient retrofits.

### 5.3. Computational Advantages of Statistical Learning Methods

The previous sections in this paper have analyzed and discussed the accuracy of the proposed Random Forest models, which is typically the concern for surrogate models. This section briefly discusses the advantages for the use of the Random Forest models in terms of the efficiency in computation time for estimating network performance for future data points. For the hydraulic simulation strategy, the direct computation time is reported. For the statistical learning strategies, the computation time for data assembly, metric calculations, model training, and model prediction are reported. Results are tabulated in Table 2, and reflect empirical computation times performed in series (i.e. omitting opportunities for parallel processing) using a 64 bit Windows 10 laptop with an Intel(R) Core(TM) i7-5500 2.40 GHz processor.

Table 2 shows that the primary cost of the Naïve strategy is model training, while most of the time for the Metric strategies is spent calculating the explanatory metrics. The Random Forest models significantly reduce computational expense: in comparison to hydraulic simulation, its computation time is 1/30th with the Naïve strategy and 1/600th with the Metric strategy. Comparing the Random Forest models, the training times of the Metric strategies are significantly better than that of the Naïve strategy, and outweigh the time required to compute the explanatory metrics. This is due to the Naïve model having to grow decision trees using a pool of 6,307 predictor variables, while the Metric strategies only deal with 49 for the Full Set and 3 for the Optimized Set. Further, predictions are effectively instant for all Random Forest models. Thus,

once the model is constructed, subsequent network performance estimation is very fast, facilitating large scale estimation and rapid assessment.

One disadvantage for using Random Forest models not reflected in Table 2 is that additional time must be spent on precalculations and training of the model. Depending on the objective of the analyst, this training time may or may not be a critical consideration. Two situations where training time is not critical are as follows. First, for rapid network assessment of recovery strategies following a catastrophic event, a Random Forest model can be prepared beforehand to enable rapid post-event predictions. Second, when considering retrofit strategies, initial training can be done once, and then used when evaluating several possible strategies.

## 6. Conclusion

This paper demonstrated the application of the Random Forest statistical learning technique for estimating the reliability of water supply networks subject to seismic damage. Two strategies were explored: a Naïve strategy, in which network component damage states are used as predictor variables, and a Metric strategy, in which network component states are translated to a set of explanatory metrics, which are used as predictor variables. Twenty-three metrics were considered, yielding 49 potential predictor variables when considering their application to specific portions of the network. An optimized set of three predictor variables were identified using the out-of-bag permuted variable delta error as the measure of variable importance. The Random Forest model using the optimized set was seen to perform nearly as well as with the full set of explanatory variables while performing better than the Naïve strategy. Additionally, the effect of training data size on predictive power was assessed.

The retrofit strategies identified using Random Forest predictions performed similarly to those determined using hydraulic simulation. Then, the sensitivity of retrofit effectiveness to the training data size and predictive power of the statistical model was considered, and the retrofit identification approach was seen to be robust down to a certain minimum size of training data. The Random Forest strategies offer great computation time advantages relative to direct hydraulic simulation, when the time to pre-train such models can be afforded. These results indicate that statistical learning methods are powerful candidates for rapid network reliability assessment, especially for large networks.

Future work could refine the hyperparameters for each Random Forest model, and explore methods for developing training data that better covers the parameter space. Another opportunity for research is to assess the proposed methodology on other networks. Though the predictive power and optimized metrics reported from this case study are specific to this network, we anticipate that the approach should be broadly applicable.

This paper has demonstrated the potential of statistical models to supplement or replace full physical characterization and simulation for the estimation of network performance. Moreover, Random Forest models, in combination with a set of explanatory metrics, have been shown to adequately replicate the network performance estimates and retrofit strategies identified using hydraulic simulation. These features make it an attractive candidate for infrastructure network risk assessment and decision support.

## 7. Acknowledgments

This work was supported in part by the National Science Foundation under NSF grant number CMMI 09524020. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank Charles Scawthorn, David Myerson, Aaron Lee, and others at the San Francisco Public Utilities Commission and AECOM for their data and technical discussions. The authors also acknowledge the support provided by Mohammad Javanbarg, Thomas O’Rourke, Mahalia Miller, Christophe Loth, Camilo Gómez, and Abhineet Gupta.

## 8. References

- [1] J. Wu, J. Baker, End-to-end simulation and analysis framework for efficient seismic retrofitting of water systems, in: 16th World Conference on Earthquake Engineering, Santiago, Chile, 2017.  
URL [http://web.stanford.edu/~bakerjw/Publications/Wu\\_Baker\\_\(2017\)\\_Framework,\\_16WCEE.pdf](http://web.stanford.edu/~bakerjw/Publications/Wu_Baker_(2017)_Framework,_16WCEE.pdf)
- [2] J. Wu, End-to-end seismic risk analysis framework for the identification of infrastructure network retrofits, Ph.D. thesis, Stanford University, Stanford, CA (Aug. 2017).
- [3] F. Meng, G. Fu, R. Farmani, C. Sweetapple, D. Butler, Topological attributes of network resilience: a study in water distribution systems, *Water Research* 143 (2018) 376–386.
- [4] R. Wright, M. Herrera, P. Parpas, I. Stoianov, Hydraulic resilience index for the critical link analysis of multi-feed water distribution networks, *Procedia Engineering* 119 (2015) 1249–1258.
- [5] R. Dzedzic, B. Karney, Water distribution system performance metrics, *Procedia Engineering* 89 (2014) 363–369.
- [6] W. Liu, Z. Song, Review of studies on the resilience of urban critical infrastructure networks, *Reliability Engineering and System Safety* 193.
- [7] Q. Shuang, H. Liu, E. Porse, Review of the Quantitative Resilience Methods in Water Distribution Networks, *Water* 11 (2019) 1189.
- [8] P. Juan-Garcia, D. Butler, J. Comas, G. Darch, C. Sweetapple, A. Thornton, L. Corominas, Resilience theory incorporated into urban wastewater systems management. State of the art, *Water Research* 115 (2017) 149–161.
- [9] A. A. Ganin, E. Massaro, A. Gutfraind, N. Steen, J. M. Keisler, A. Kott, R. Mangoubi, I. Linkov, Operational resilience: concepts, design and analysis, *Scientific Reports* 6 (2016) 19540.  
URL <https://doi.org/10.1038/srep19540>
- [10] A. Cuppens, I. Smets, G. Wyseure, Definition of realistic disturbances as a crucial step during the assessment of resilience of natural wastewater treatment systems, *Water Science and Technology* 65 (8) (2012) 1506–1513.
- [11] M. Ouyang, L. Duenas-Osorio, Time-dependent resilience assessment and improvement of urban infrastructure systems, *Chaos* 22.
- [12] A. Assad, O. Moselhi, T. Zayed, A new metric for assessing resilience of water distribution networks, *Water* 11 (2019) 1701.
- [13] M. Herrera, E. Abraham, I. Stoianov, A graph-theoretic framework for assessing the resilience of sectorised water distribution networks, *Water Resource Management* 30 (2016) 1685–1699.
- [14] J. J. Bommer, H. Crowley, The Influence of Ground-Motion Variability in Earthquake Loss Modelling, *Bulletin of Earthquake Engineering* 4 (3) (2006) 231–248. doi:10.1007/s10518-006-9008-z.  
URL <http://link.springer.com/article/10.1007/s10518-006-9008-z>
- [15] J. Wu, L. Dueñas Osorio, Calibration and validation of a seismic damage propagation model for interdependent infrastructure systems, *Earthquake Spectra* 29 (3) (2013) 1021–1041.  
URL <http://www.earthquakespectra.org/doi/abs/10.1193/1.4000160>
- [16] M. Fragiadakis, S. E. Christodoulou, Seismic reliability assessment of urban water networks, *Earthquake Engineering & Structural Dynamics* 43 (3) (2014) 357–374. doi:10.1002/eqe.2348.  
URL <http://doi.wiley.com/10.1002/eqe.2348>
- [17] S. E. Christodoulou, M. Fragiadakis, Vulnerability Assessment of Water Distribution Networks Considering Performance Data, *Journal of Infrastructure Systems* 21 (2) (2015) 04014040. doi:10.1061/(ASCE)IS.1943-555X.0000224.  
URL <http://ascelibrary.org/doi/10.1061/%28ASCE%29IS.1943-555X.0000224>
- [18] S. Esposito, I. Iervolino, A. d’Onofrio, A. Santo, F. Cavalieri, P. Franchin, Simulation-Based Seismic Risk Assessment of Gas Distribution Networks: Seismic risk assessment of gas networks, *Computer-Aided Civil and Infrastructure Engineering* 30 (7) (2015) 508–523. doi:10.1111/mice.12105.  
URL <http://doi.wiley.com/10.1111/mice.12105>
- [19] A. Pandit, J. C. Crittenden, Index of network resilience (INR) for urban water distribution, 2012.
- [20] D. Soldi, A. Candelieri, F. Archetti, Resilience and vulnerability in urban water distribution networks through network theory and hydraulic simulation, *Procedia Engineering* 119 (2015) 1259–1268.
- [21] Y. Setiadi, T. Tanyimboh, A. Templeman, Modelling errors, entropy and the hydraulic reliability of water distribution systems, *Advances in Engineering Software* 36 (11-12) (2005) 780–788. doi:10.1016/j.advengsoft.2005.03.028.  
URL <https://linkinghub.elsevier.com/retrieve/pii/S0965997805001274>
- [22] A. Pagano, I. Pluchinotta, R. Giordano, M. Vurro, Drinking water supply in resilient cities: Notes from L’Aquila earthquake case study, *Sustainable Cities and Society* 28 (2017) 435–449.
- [23] Z. Xu, J. E. Ramirez-Marquez, Y. Liu, T. Xiahou, A new resilience-based component importance measure for multi-state networks, *Reliability Engineering and System Safety* 193.
- [24] G. Galvan, J. Agarwal, Assessing the vulnerability of infrastructure networks based on distribution measures, *Reliability Engineering and System Safety* 196 (2020) 106743. doi:<https://doi.org/10.1016/j.ress.2019.106743>.
- [25] M. Maiolo, D. Pantusa, M. Carini, G. Capano, F. Chiaravalloti, A. Procopio, A new vulnerability measure for water distribution network, *Water* 10 (2018) 1005.
- [26] P. Hines, E. Cotilla-Sanchez, S. Blumsack, Do topological models provide good information about electricity infrastructure vulnerability?, *Chaos: An Interdisciplinary Journal of Nonlinear Science* 20 (3) (2010) 033122. doi:10.1063/1.3489887.  
URL <http://scitation.aip.org/content/aip/journal/chaos/20/3/10.1063/1.3489887>
- [27] S. D. Guikema, Natural disaster risk analysis for critical infrastructure systems: An approach based on statistical learning theory, *Reliability Engineering & System Safety* 94 (4) (2009) 855–860. doi:10.1016/j.ress.2008.09.003.  
URL <http://linkinghub.elsevier.com/retrieve/pii/S0951832008002317>

- [28] G. Perrin, Active learning surrogate models for the conception of systems with multiple failure modes, *Reliability Engineering & System Safety* 149 (2016) 130–136. doi:10.1016/j.ress.2015.12.017.  
URL <http://linkinghub.elsevier.com/retrieve/pii/S0951832015003695>
- [29] S.-R. Han, S. D. Guikema, S. M. Quiring, K.-H. Lee, D. Rosowsky, R. A. Davidson, Estimating the spatial distribution of power outages during hurricanes in the Gulf coast region, *Reliability Engineering & System Safety* 94 (2) (2009) 199–210. doi:10.1016/j.ress.2008.02.018.  
URL <http://linkinghub.elsevier.com/retrieve/pii/S0951832008000665>
- [30] K. Rokneddin, J. Ghosh, L. Dueñas Osorio, J. Padgett, Seismic reliability assessment of bridge networks by statistical learning, in: 11th International Conference on Structural Safety and Reliability, New York, NY, 2013.  
URL [https://www.researchgate.net/publication/281179055\\_Seismic\\_reliability\\_assessment\\_of\\_bridge\\_networks\\_by\\_statistical\\_learning](https://www.researchgate.net/publication/281179055_Seismic_reliability_assessment_of_bridge_networks_by_statistical_learning)
- [31] A. Robles-Velasco, P. Cortes, J. Muuzuri, L. Onieva, Prediction of pipe failures in water supply networks using logistic regression and support vector classification, *Reliability Engineering and System Safety* 196 (2020) 106754. doi:https://doi.org/10.1016/j.ress.2019.106754.
- [32] J. M. Hernandez, P. Van Mieghem, Classification of graph metrics, Delft University of Technology, Tech. Rep.  
URL [https://www.nas.ewi.tudelft.nl/people/Piet/papers/TUDreport20111111\\_MetricList.pdf](https://www.nas.ewi.tudelft.nl/people/Piet/papers/TUDreport20111111_MetricList.pdf)
- [33] J. Li, L. Dueñas Osorio, C. Chen, B. Berryhill, A. Yazdani, Characterizing the topological and controllability features of U.S. power transmission networks, *Physica A: Statistical Mechanics and its Applications* 453 (2016) 84–98. doi:10.1016/j.physa.2016.01.087.  
URL <http://linkinghub.elsevier.com/retrieve/pii/S0378437116001515>
- [34] J. M. Torres, L. Duenas-Osorio, Q. Li, A. Yazdani, Exploring Topological Effects on Water Distribution System Performance Using Graph Theory and Statistical Models, *Journal of Water Resources Planning and Management* (2016) 04016068doi:10.1061/(ASCE)WR.1943-5452.0000709.  
URL <http://ascelibrary.org/doi/10.1061/%28ASCE%29WR.1943-5452.0000709>
- [35] S. Dunn, S. M. Wilkinson, Identifying Critical Components in Infrastructure Networks Using Network Topology, *Journal of Infrastructure Systems* 19 (2) (2013) 157–165. doi:10.1061/(ASCE)IS.1943-555X.0000120.  
URL <http://ascelibrary.org/doi/10.1061/%28ASCE%29IS.1943-555X.0000120>
- [36] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, Springer, 2003.  
URL [http://www.newbooks-services.de/MediaFiles/Texts/0/9780387848570\\_Intro\\_001.pdf](http://www.newbooks-services.de/MediaFiles/Texts/0/9780387848570_Intro_001.pdf)
- [37] L. Breiman, Random forests, *Machine learning* 45 (1) (2001) 5–32.  
URL <http://link.springer.com/article/10.1023/A:1010933404324>
- [38] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems, *Journal of Machine Learning Research* 15 (1) (2014) 3133–3181.  
URL <http://www.jmlr.org/papers/volume15/delgado14a/source/delgado14a.pdf>
- [39] Mathworks, Create ensemble of bagged decision trees - MATLAB (May 2017).  
URL <https://www.mathworks.com/help/releases/R2013a/stats/treebagger.html?searchHighlight=random%20forest>
- [40] C. Scawthorn, Analysis of Fire Following Earthquake Potential for San Francisco, California, Tech. rep., SPA Risk LLC, Berkeley, CA (Jul. 2010).
- [41] D. Myerson, A. C. Symonds, Planning Support Services for San Francisco’s Auxiliary Water Supply System (AWSS), in: *Pipelines 2014: From Underground to the Forefront of Innovation and Sustainability*, 2014, pp. 1784–1793.  
URL <http://ascelibrary.org/doi/abs/10.1061/9780784413692.162>
- [42] SFPUC, Emergency Firefighting Water System (2018).  
URL <https://sfwater.org>
- [43] E. H. Field, T. E. Dawson, K. R. Felzer, A. D. Frankel, V. Gupta, T. H. Jordan, T. Parsons, M. D. Petersen, R. S. Stein, R. J. Weldon, C. J. Wills, Uniform California Earthquake Rupture Forecast, Version 2 (UCERF 2), *Bulletin of the Seismological Society of America* 99 (4) (2009) 2053–2107. doi:10.1785/0120080049<p>.  
URL <http://www.bssaonline.org/cgi/content/abstract/99/4/2053>
- [44] D. M. Boore, G. M. Atkinson, Ground-Motion Prediction Equations for the Average Horizontal Component of PGA, PGV, and 5%-Damped PSA at Spectral Periods between 0.01 s and 10.0 s, *Earthquake Spectra* 24 (1) (2008) 99–138. doi:10.1193/1.2830434.  
URL <http://earthquakespectra.org/doi/abs/10.1193/1.2830434>
- [45] M. Miller, Seismic risk assessment of complex transportation networks, Ph.D. thesis, Stanford University (2014).
- [46] S.-S. Jeon, T. D. O’Rourke, Northridge Earthquake Effects on Pipelines and Residential Buildings, *Bulletin of the Seismological Society of America* 95 (1) (2005) 294–318. doi:10.1785/0120040020.  
URL <http://www.bssaonline.org/cgi/doi/10.1785/0120040020>
- [47] AECOM, AGS, CS-199 Planning Support Services for Auxiliary Water Supply System (AWSS) - Technical Memorandum (TM) - Task 8 - Modeling (Jan. 2012).
- [48] L. A. Rossman, EPANET 2 Users Manual, United States Environmental Protection Agency, Cincinnati, OH, 2000.
- [49] Cornell University, GIRAFFE User’s Manual 4.2 (2008).
- [50] A. Yazdani, P. Jeffrey, Complex network analysis of water distribution systems, *Chaos: An Interdisciplinary Journal of Nonlinear Science* 21 (1) (2011) 016111.  
URL <http://scitation.aip.org/content/aip/journal/chaos/21/1/10.1063/1.3540339>